Academic vocabulary in Wikipedia articles : Frequency and dispersion in uneven datasets

Hiltunen, Turo

Brill
2019

Hiltunen , T & Tyrkkö , J 2019 , Academic vocabulary in Wikipedia articles : Frequency and dispersion in uneven datasets . in C Suhr , T Nevalainen & I Taavitsainen (eds) , From data to evidence in English Language research . Language and computers: Studies in digital þÿ l i n g u i s t i c s , n o . 8 3 , B r i l l , L e i d e n , p p . 2 8 2 3 0 6 . h t t p s : / / d o i . o r g / 1 0 . 1

http://hdl.handle.net/10138/325621
https://doi.org/10.1163/9789004390652_013

unspecified
acceptedVersion

Academic vocabulary in Wikipedia articles: Frequency and dispersion in uneven datasets

1 Introduction

In a decade and a half since its launch in 2001, Wikipedia has become the largest and most widely used encyclopaedia in the world. Currently, the English-language Wikipedia comprises more than 4.75 million articles. Organised on the principles of shared authorship and crowdsourcing, Wikipedia benefits from contributions by more than 75,000 registered volunteers and countless occasional editors.

Despite such widespread popularity, the status of Wikipedia in higher education (HE) settings remains somewhat controversial. Numerous concerns have been expressed over the years about factual inaccuracies, bias and lack of stability of articles (e.g. Kamm 2007, Myers 2010).[1] Stylistically, Wikipedia articles also show considerable variation, which is hardly surprising due to the range and variety of topics covered as well as the heterogeneous background of the editors and contributors (Myers 2010; Hiltunen 2014). Rosenzweig (2006) criticises this very aspect of Wikipedia articles on history topics. For example, he notes that although the Wikipedia account of Abraham Lincoln is factually correct, it is stylistically "verbose and dull", devoid of clarity and engagement expected of good historical writing.

This being the case, the Wikipedia article — which Kuteeva (2011: 46) describes as a "new academic genre" —is likely to be less suitable as a model of academic style, which students in higher education are required to master, than published expository texts, including traditional encyclopaedia texts written and edited by experts. At the same time, while corpus-based descriptions of traditional academic genres (e.g. research articles, review articles and theses) are ubiquitous, similar accounts are largely lacking for Wikipedia articles, which makes systematic comparisons difficult. More empirical research is therefore clearly needed to describe the characteristics of Wikipedia articles from a linguistic, stylistic and rhetorical point of view. In this article, we treat Wikipedia

[1] See https://en.wikipedia.org/wiki/Wikipedia:Criticisms for more quotations from critics of Wikipedia.

articles and research articles as different genres due to their different communicative purposes (see e.g. Swales 1990), and compare their linguistic characteristics using corpus data. Communicative purpose is the main criterion for identifying genres especially in English for Academic purposes (EAP) settings, though not the only one (see e.g. Hiltunen 2010: 27–30), and the definition also allows variation between individual instances of a genre (Swales 1990: 49).[2]

One key aspect of academic style concerns the use of academic vocabulary. Academic texts contain large numbers of lexical items that are not frequently used in non-academic contexts, yet knowing them is essential for understanding these texts (Schmitt et al 2011, Davies & Gardner 2014). Acquiring this vocabulary, which includes both general academic vocabulary and content-specific terminology, typically requires specific instruction, especially in ESL/EFL contexts (Hiebert & Lubliner 2008: 107). The use of specific vocabulary items is certainly not the only defining characteristic of academic texts; in addition, they differ from other registers also grammatically, as shown in Biber et al.'s (1999) corpus-based descriptive grammar. However, in this paper we focus specifically on academic vocabulary, arguing that a corpus-based comparison of use frequency of this lexical field serves as a useful indicator of genre. In this exploratory paper, we take a data-driven approach to assessing the use of academic vocabulary in Wikipedia articles by selecting a representative sample of articles from three different academic disciplines and comparing them to published research articles from the same disciplines. Our analysis is based on a set of academic words, the well-known Academic Word List (AWL) compiled by Coxhead (2000). Our data comes from the Westbury Lab Wikipedia Corpus (see Shaoul and Westbury 2010) of slightly less than 2 million articles, which we contrast with a comparable set of academic research articles, which have been peer-reviewed and published (see Hiltunen 2010, Hiltunen & Mäkinen 2014). In our analysis, we employ methods of statistical data analysis to classify samples from the corpus according to the frequencies of AWL words.[3] The unsupervised classification procedure will group the articles according to academic content regardless of topic, which allows us to measure genre-specific similarities. We also address some methodological issues with this type of data, including the treatment of texts of different length.

[2] Kuteeva's (2011) treatment of the Wikipedia article as an academic genre relies on Myers' (2010) analysis, which follows Berkenkotter and Huckin's (1995) sociocognitive definition of genres as rhetorical structures that can be manipulated according to the situation.

[3] Unless otherwise specified, the term "academic word" refers to a word appearing on Coxhead's (2000) Academic Word List.

The findings of the study show that AWL words are common in both genres in focus, and more interestingly, if we look at their aggregate frequencies, Wikipedia articles are not markedly different from RAs within the same discipline. At the same time, we can observe disciplinary differences in the distribution of AWL words in Wikipedia, as can be expected, such that Economics writing contains more tokens than the other two disciplines. Disciplinary differences can likewise be observed in the distribution of individual words.

2 Background

2.1 Previous work on Wikipedia

As Wikipedia has become an increasingly important resource of knowledge in the world at large, it has also attracted a great deal of scholarly interest across disciplines. Wikipedia's own, incomplete list of studies on different aspects of the online encyclopaedia currently includes over 1,500 peer-reviewed journal articles and 4,000 academic conference presentations.[4] A Wikipedia article entitled Academic studies about Wikipedia suggests that these studies fall under two major categories: those analysing the production and reliability of content, and those investigating social aspects of knowledge generation.[5] Linguistic studies on Wikipedia are also numerous, and the majority of these use Wikipedia text for different types of Natural Language Processing (NLP) applications (for an overview, see Medeleyan et al 2009), mainly due to the volume of diverse texts available. Descriptive linguistic analyses are much fewer in comparison, as are discourse-analytical studies on Wikipedia (for an overview, see Myers 2010 and Hiltunen 2014).

From the perspective of this paper, it is interesting to note that the use of Wikipedia in higher education has recently generated a number of studies exploring the question of how best to make use of this resource. Much of this research is in fact positive, highlighting the possibilities and affordances of Wikipedia and wikis for the teaching of academic literacy and academic writing (e.g. Barton & Cummings eds. 2008, Tardy 2010, Miller 2012,). Kuteeva, for instance, has suggested that using wikis may help students of English for Academic Purposes (EAP) courses acquire a wider sense of audience than traditional teaching methods and provide increased opportunities for

[4] The list is available at: http://wikipapers.referata.com/wiki/List_of_journal_articles (accessed 11 March 2016).

[5] https://en.wikipedia.org/wiki/Academic_studies_about_Wikipedia (accessed 11 March 2016).

collaboration, leading improved grammatical correctness and text organization (2011: 55).

Alongside these opinions, there are critical views which cast doubt on the aptness of Wikipedia in HE settings (e.g. Rosenzweig 2006, Waters 2007), despite the fact that both students and academics make frequent use of Wikipedia. Myers (2010: 143–4) has suggested that such critical opinions can even be "a sort of gut response, without much argument or experience with wikis", but systematic attempts to estimate the reliability of Wikipedia sections have also been attempted.[6] If we want to focus on style of writing and determine how Wikipedia articles are different from traditional academic genres in this respect, a nubmer of requirements need to be met: first, we clearly need samples representing both kinds of writing that are sufficiently large. In addition, we need to decide what aspect of style to focus on, and how to operationalise for corpus linguistic anlaysis that in such a way that would enable us to carry out meaningful comparisons between the samples. As our point of comparison is academic discourse, we investigate usage of a group of words that are closely associated with it, namely academic vocabulary. While wholly unexplored in previous research, this aspect of Wikipedia writing is highly relevant to determining its status in relation to the writing style of established academic genres.

2.2 Academic word list

One visible characteristic of academic prose is the kind of words that are used in texts. Within this part of vocabulary, a number of distinct categories can be identified: Nation (2001) distinguishes between, firstly, canonical technical terms (technical vocabulary) associated with the subject area discussed in the text, and secondly, words that are not specific to any particular area of inquiry but which instead occur across a wide range of academic texts — this component is known in lexical studies as academic vocabulary, and is the main focus of this paper. It should be noted that neither of these types cover all word tokens in academic texts; instead what makes up the majority are general high-frequency words, which are also common in non-academic texts.

To illustrate the differences between these three types of vocabulary, consider the following brief extract from an introduction of a medical RA included in our reference corpus, reproduced as example (1) below. We have highlighted a number of words in the quotation: italicised words are technical terms in the field of medicine, and bold type corresponds to academic words, which are common in academic texts across the board.

---

[6] See https://en.wikipedia.org/wiki/Reliability_of_Wikipedia for more information.

(1) Prolotherapy is a treatment for chronic nonspecific low-back pain that involves a protocol of ligament injections exercises and vitamin and mineral supplements. It is based on the premise that back pain results from weakened ligaments and that these ligaments can be strengthened by the injection into them of irritant proliferant solutions. These solutions variously contain phenol glycerine or hypertonic glucose mixed with local anesthetic and aim to induce inflammation and deposition of collagen fibers in the weak ligaments. There is limited histologic evidence of thickening of sacroiliac ligaments in association with a reduction in low-back pain scores and increased lumbar range of motion using all these solutions combined. The supplementary regimen of exercises and oral vitamins and minerals ostensibly promote collagen growth to induce optimal strengthening of the treated ligaments. The anecdotal and experimental evidence are contradictory. Testimonies to the effectiveness of prolotherapy include one from the former Surgeon General of the United States. (Yelland et al. 2004: 9)

High-frequency words in the extract include universally useful function words such as is, a, for, that, and and, as well as frequent lexical words such as low, back, exercises, and results. The italicised technical terms are likewise easily identified; the words prolotherapy, chronic, histological, ligament, collagen, and injection refer to the subject matter of the extract, but are clearly infrequent outside the (bio-)medical context. Other lexical items which are not among the 2,000 most frequent English words include nonspecific, ostensibly, thickening, optimal, and anecdotal, which are not specifically medical or even academic. However, our focus in this chapter is on academic words (highlighted in bold), of which there are nine instances in the extract: involves, protocol, supplements, induce, evidence, range, supplementary, promote, and contradictory. These words are commonly found across a wide range of academic texts, but they are not specific to any particular area of inquiry (Coxhead 2000: 214, 221).

Given that academic vocabulary has been identified as a challenging area for learners and novice writers (Coxhead 2000: 235), it is not surprising that the study of this component of academic writing has been primarily motivated by pedagogical concerns. Corpus-based studies have accordingly produced lists of vocabulary items, which would help learners to build their repertoire and develop their skills as writers of academic texts (for an overview, see Martinez & Schmitt 2015). Probably the best known and most widely used of these lists to date is Averil Coxhead's Academic Word List (AWL), which we also use in the present study. Coxhead describes academic words as "salient" and "supportive but not central to the topics of the texts in which they occur" (2000: 214), and her list includes some 3,000 words divided into 570 word families, which accounted for approximately 10% of the tokens in the test corpus of academic

texts.[7] Seventeen years on, the AWL is still relevant, although the list has also received some criticism for an uneven representativeness of fields (Hyland & Tse 200) as well as the using word families as an organising principle and excluding the words in West's (1953) General Service List (Davies and Gardner 2013). Recent work in corpus-based vocabulary studies have used increasingly sophisticated methods for creating word lists, addressing in particular the issue of disperson (Brezina & Gablasova 2013, Miller & Biber 2015). However, given that the coverage of the AWL has been shown to be consistent across a range of academic texts (Coxhead 2011: 356), the list offers a potentially useful benchmark for determining the degree to which the lexis of a given text collection can be treated as being "academic". At the same time, we want to emphasise that the use of AWL words is one measure of this among many; we do not claim that the presence of academic words (whether operationalised using the AWL or in some other way) would be the only relevant characteristic of an academic text, nor that the absence of these words would automatically define a text as "non-academic". On the contrary, it is well known that academic writing can be characterised with reference to a number of other features, including the frequency of specific grammatical structures (e.g. Biber 1988, 2006), preferences of co-occurrence of grammatical constructions (e.g. Hiltunen 2010) and the preferred rhetorical strategies and "move structures" (e.g. Swales 1990). We shall return to this issue in Section 6.

3 Material

In recent years, a number of different Wikipedia corpora have been made available for corpus linguistic research. This study draws on Westbury Lab Wikipedia Corpus (WLWC, Shaoul & Westbury 2010), a 990-million word corpus based on a snapshot of the English-language Wikipedia from April, 2010.[8] From this source data we extracted a smaller corpus for detailed analysis.

Because the WLWC is released as a plain text corpus without information about the linkedness of articles, the samples were extracted based on article titles. Our

[7] The full list of AWL items is found in Appendix A of Coxhead (2000) and is available online at http://www.victoria.ac.nz/lals/resources/academicwordlist/.

[8] Other recent Wikipedia corpora include the Wikipedia XML corpus (Denoyer and Gallinari 2006), Wikicorpus (Reese et al. 2010), WaCkypedia_EN, (Baroni et al. 2009), and Wikipedia Talk Page Conversations Corpus (Danescu-Niculescu-Mizil et al 2012.).

three subcorpora consist of all the articles linked to from the three main pages; to obtain the list of linked articles, we consulted the original Wikipedia pages outside the WLWC. We are particularly interested in exploring variation across different disciplines, given that in previous studies discipline has emerged as one of the main factors accounting for variation within academic prose (e.g. Hyland 2000). To what extent this applies to Wikipedia articles is still an open question. Accordingly, we extracted three moderately sized subcorpora from Wikipedia, each of which represents one area of inquiry: economics, medicine, and literary criticism (see also Hiltunen 2014, forthcoming). These subcorpora, and the academic disciplines they represent, belong to different "disciplinary cultures", which influence the writers approach their subject matter and write about it (see Becher & Trowler 2001).

As previously mentioned, the length of texts varies considerably in Wikipedia, but the right panel in Figure 1 shows that texts of different length are distributed fairly evenly across the three subcorpora. We use research articles as a benchmark in our analysis. For both medicine and literary criticism, we analysed the collections of 64 articles used previously in Hiltunen (2010), and for economics, a collection of 50 articles described in Hiltunen & Mäkinen (2014: 351–3) was used. RAs display more variation in text length and the median word count is considerably larger than in Wikipedia articles (Figure 1, right panel).

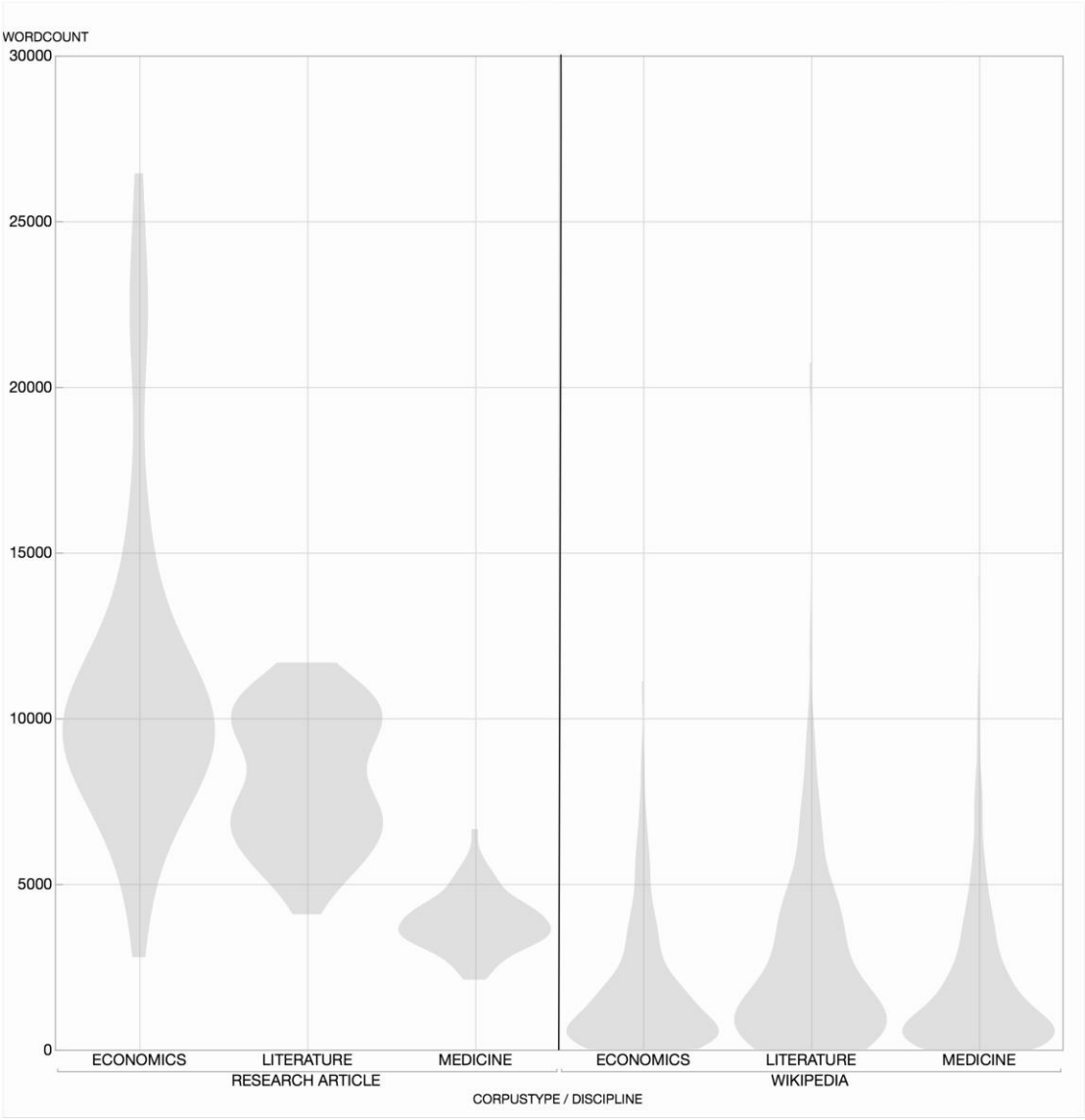


Figure 1. A violin plot of article length (measured as tokens per text) across genres and disciplines.

Table 1. Article word counts and word count dispersion across genres and disciplines

Level	n	Word count	Mean	SD	Median
Journal_econ	50	559,000	11,182	4,782	9,991

Journal_lit	64	524,224	8,191	1,994	8,015
Journal_med	64	248,064	3,876	875	4,363
Wiki_econ	470	855,870	1,821	1,935	1,150
Wiki_lit	182	407,134	2,237	2,737	1,565
Wiki_med	439	856,050	1,950	2,125	1,107

4 Methods

4.1 How to assess the frequency of AWs?

Determining and quantifying the frequencies of linguistic features can be done in a variety of ways, and the suitability of the approach depends on the research goals at hand. It is necessary to briefly consider the relative merits of different approaches, given that we are dealing with corpora that are different from many commonly used corpora in more than one respect: Wikipedia articles are collectively authored and edited, and their length and quality varies greatly, which in turn raises the issues of comparability and representativeness. To give a sense of the editing history of important Wikipedia articles, the article on "Adam Smith", the Scottish philosopher and economist, was created on Sep 10, 2001. By April 2010, the time the WLWC corpus was compiled, the article had been edited 4,856 times by hundreds of different editors.

In many studies, corpora are simply approached as monolithic entities without regard for the dispersion of the phenomena of interest within the corpus. In this approach, which has come to be known as the bag-of-words model (Manning & Schütze 1999: 237; Evert 2005), the frequencies of linguistic features are calculated using the total word count of the corpus and, more often than not, standardized using whatever base seems appropriate. In the best case scenario, in which the feature of interest is more or less evenly distributed across the different texts that make up the corpus, the bag-of words method can provide a relatively realistic view of the phenomenon at hand. However, the bag-of-words method may lead to significant misrepresentation of the true population frequencies, particularly when it comes to lexical phenomena (Gries 2008, 2009).

It can be argued that the issue of dispersion is less serious in a large randomly sampled corpus, where the random sampling would ensure that the corpus provides a realistic overall representation of the language or register under investigation. However, with small corpora where the sampling method is something other than random, the bag-of-words approach is problematic (see, e.g., Evert 2005 and Kilgarriff 2005 for discussion). Therefore, despite the fact that this is the dominant approach,[9] these distributional assumptions cannot necessarily be taken for granted in our data, given that we are dealing with two genres with very different characteristics. While many grammatical and phraseological features are indeed reasonably evenly distributed across individual texts (Biber 1993), topic-related lexical phenomena depend entirely on which texts happened to be included in the corpus. For example, the

[9] For example, Gries (2009: 198) observes that corpus-linguistic studies attempting to take account of dispersion in the analysis of data, either by quantifying the homogeneity of distributions or using adjusted frequencies, are in the minority.

specialised terminology of a given profession is only likely to occur in texts belonging to that professional community, and if a small or medium-sized corpus happens to include several such texts, the specialised terms may show artificially inflated frequencies. Miller and Biber (2015) have recently shown that even highly restricted discourse domains like undergraduate psychology textbooks display remarkable amounts of lexical variation. To safeguard against these problems, different measures of dispersion should be incorporated into the analytical design.

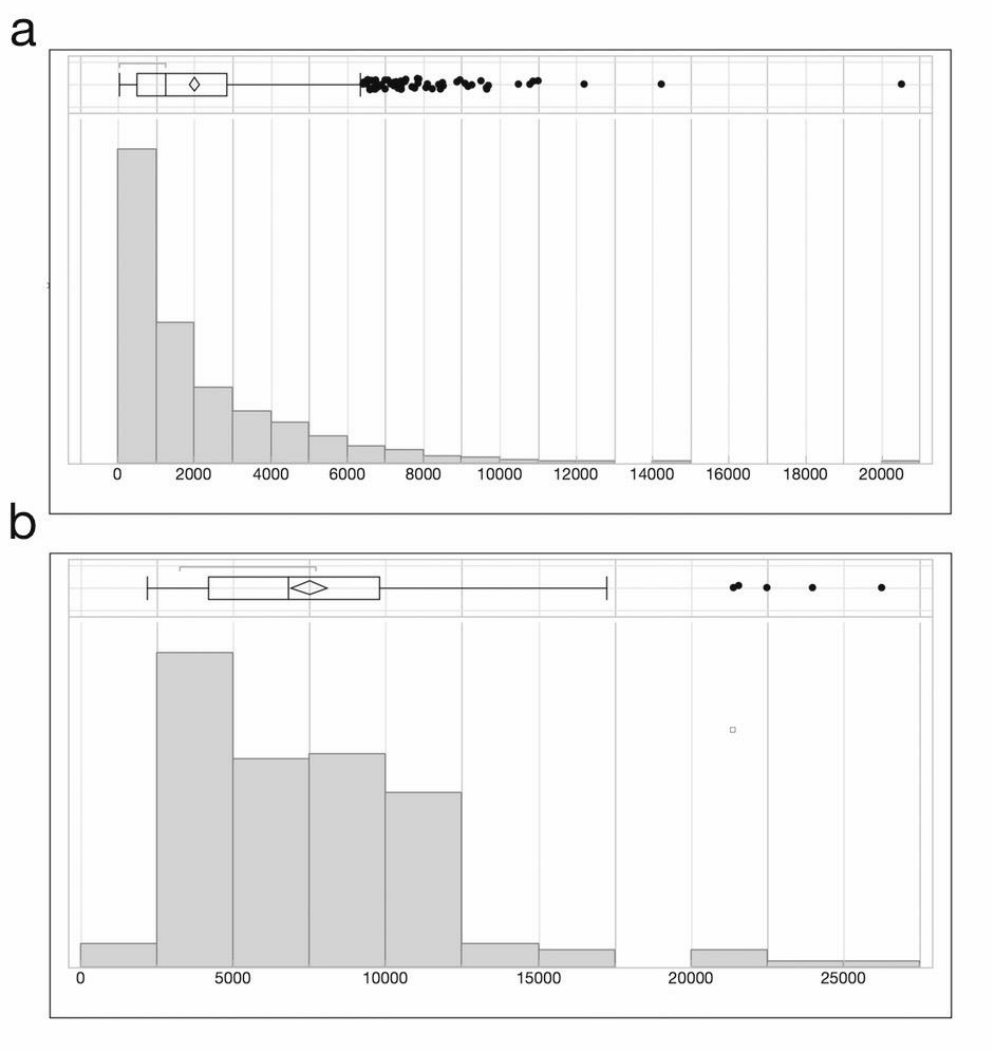
When dispersion is taken into account in corpus linguistics, the most common method is to treat each individual text as an independent observation and to calculate means and dispersion metrics based on standardized frequencies in each observation (e.g. Biber and Jones 2009). Although the overall frequency of the phenomenon of interest is not affected by this method, after all both the overall number of hits and the total word count remain the same as in the bag-of-words approach, dispersion will be reflected in a dispersion metric such as sample standard deviation.[10]

All of the methods discussed so far are predicated on the assumption that the individual texts are what one might intuitively describe as reasonably long extracts or full texts. That sounds very vague, but as surprising as it may be, there is not much explicit discussion in corpus linguistic literature on how long individual samples should be, or what the effects of sample length might be on quantitative analysis. There are some well-known rules-of-thumb, such as 2,000-word extracts being sufficiently long for studying common grammatical features (Biber 1993, Ide et al. 2002), or 10,000–20,000 word extracts being long enough for medium frequency lexical phenomena (see Kennedy 1998, Nelson 2010: 58). but relatively little attention is paid to what happens when the sizes of the basic unit of observation, typically texts, are either significantly variant or very short. As it happens, although standardized frequency is very a useful and necessary metric in corpus linguistics, it is easily skewed when the texts get very short. In the worst case, scholars sometimes standardize frequency data to a base that is greater in value than the text being examined — for example, standardizing the frequency found in a 2,000-word long text to a base of 10,000 words — which effectively means extrapolating up, or making the claim that were the text longer, the occurrences found would consistently scale up. The absurdity of the systematic error in this scenario becomes evident if we imagine what happens with extremely short samples. If we took a random 100-word extract from a

[10] Standard deviation is a common metric for measuring variation in a data set. Population standard deviation is used when the data set used is the object of examination, while the adjusted metric of sample standard deviation is used when the data set is a sample from a larger population. Corpora are typically the latter.

novel and found 10 occurrences of word A, would it make sense to claim that the frequency of word A is 100/10,000 words in novels? Of course not.

The Wikipedia data examined in the present study makes a useful case study when it comes to sample length.[11] One characteristic of the WLWC is the presence of a large number of very short texts (see Hiltunen 2014: sect. 4.1). In our sample, the median length of the Wikipedia articles is 1,220 words, with 25% of the articles being less that 500 words in length (see Figure 2). By comparison, the median length of the research articles is 6,802 words, more than 5 times longer.



[11] Many other text types within the field of computer-mediated communication, such as emails, text messages and tweets, present similar problems related to short and sometimes extremely short samples.

Figure 2. Histograms of word counts per file in Wikipedia corpus (a) and RA corpus (b).

Our solution to the problem was to examine the articles in short chunks.[12] Not only does this allow us to include the shorter Wikipedia articles in the analysis in a way that goes around the problem of skewed standardised frequencies, but this also gives a more accurate assessment of longer Wikipedia articles which, as discussed earlier, are the product of a collaborative effort of hundreds, sometimes thousands of contributions, minor edits and revisions. Looking at the data, we see that the 10% quantile of the word count in Wikipedia articles was 186 words, meaning 90% of all the articles are longer than 186 words. Rounding up, we decided to use 200 words as our chunk length. We wrote a simple script to chop all the articles into 200-word long chunks, and used those chunks as the basic unit of observation. Because we were not concerned with analysing the structures of Wikipedia articles or research articles, the chunks were treated as independent units. The objective was to assess the mean frequency and dispersion of academic words per 200-word chunk in each of the six sub-corpora. This seemed particularly useful in view of the fact that Wikipedia articles are successively authored and edited by sometimes up to hundreds of contributors, and consequently neither the internal structure nor the lexicon of an article can be taken to represent any one author's language use.

4.2 Unsupervised grouping methods

In the present study, we are primarily interested in assessing the lexical similarities and dissimilarities between Wikipedia articles and research articles as well as between articles representing different academic disciplines. Because our primary aim is to examine the use of academic words collectively rather than comparing the use of individual academic words as independent phenomena, we need a method that allows us to draw conclusions based on the usage of a large number of words at the same time. To do this, we want to observe the differences in the use of the 570 individual word families in the AWL across all the subcorpora, and to sum up these differences in a way that allows us to understand how, on the one hand, the texts relate to each other and, on the other, whether the lexical items show patterns of covariance. To answer these two questions, we turn to two statistical methods, Hierarchical Cluster Analysis and Principal Component Analysis. Both are known as unsupervised methods, because the grouping is not based on pre-existing training data that would be labelled according to group membership.

Hierarchical Cluster Analysis (HCA) is a statistical method that identifies similarities between observations (in corpus linguistics, typically texts) by

[12] This approach is similar to that used by Miller and Biber (2015) to evaluate the internal representativeness of a corpus.

analysing how similar or dissimilar they are when it comes to the values of any number of variables shared by all texts. In the present study, we used HCA to group the texts on the basis of similarities in the frequencies of the pre-selected academic words in a fully data-driven fashion. Depending on the specific clustering method selected, the texts are either split into progressively smaller clusters (divisive clustering) or joined into progressively larger clusters (agglomerative clustering). The specific distance metrics and linkage methods used have an impact on the shape, number and composition of the clusters, but more generally speaking all clustering methods produce tree diagrams or dendrograms which allow us to conceptualise the multivariate relationships between the observations in a way that would be nearly impossible using human intuition alone. In the present case, the grouping was based on the frequencies of the 570 word families, calculated as aggregate frequencies of all the items that belong to the the respective word families, as defined in Coxhead (2011).

We used Ward's method, also known as Ward's minimum variance method, in which pairs of clusters are progressively merged in a stepwise fashion based on the error sum of squares. The clustering is agglomerative in nature, starting with each individual observation (=text) as a cluster of one, and at each step the method then tests all available clusters and creates new clusters out of pairs of existing clusters in such a way that the next cluster to be created shows the minimum increase in total within-cluster variance out of all the possible new clusters.[13] Ward's method has a tendency to produce smaller clusters and more outliers than some other clustering methods.

Although it is not necessary, and occasionally not desirable, to standardize the variable values, it is a common preliminary step which prevents scale differences from skewing the clustering by giving more weight to variables with higher values. The most common method for standardizing the variables is by using z-scores, that is, by calculating the mean and standard deviation independently for each variable, and then calculating the z-scores for each observation. The z-scores are then used in the distance calculations during the clustering.

To examine the distributional tendencies of the lexical items in relation to one another, we used Principal Component Analysis (PCA). PCA is another statistical grouping method that is particularly useful when the dataset includes a great number of variables for each observation and we wish to identify underlying structures in the data by finding ways of looking at the data so that we maximize variance. This is typically conceptualised as plotting two variables on a two-dimensional coordinate system and fitting the straight line, known as an eigenvector, which produces the maximum amount of variance, or the greatest eigenvalue. If there are only two variables, there can only be two eigenvectors, of

[13] For other applications of cluster analysis in linguistics, see, e.g., Hoover (2003) and Tyrkkö (2013).

which the one that has the greatest eigenvalue is the first principal component, and the eigenvector that gives the second greatest eigenvalue is the second principal component. With only two variables, the eigenvectors are at a 90-degree angle to each other; one can think of the two vectors as forming a new set of coordinates that is based on the characteristics of the data. The more variables we have in the dataset, the more eigenvectors we can theoretically find because each new variable adds one more dimension. Admittedly, more than three dimensions can be a little difficult to conceptualise using our everyday human experience, but the additional dimensions are not mathematically difficult to compute. However, although it is possible to find as many eigenvectors as there are variables, for analytical purposes only the first two or three eigenvectors with the greatest eigenvalues are usually of interest. The eigenvectors with the greatest eigenvalues represent the most important or informative ways of looking at the dataset, while eigenvectors with small values are less informative. When principal components are reported, it is conventional to give the proportion of the overall variance explained by each component; the cumulative percentage will reach 100% when all components are included in the model. As in the HCA analysis, the AWL word families were treated as variables using the aggregate frequencies of all items included under each word family.

Consequently, we can think of PCA as a means of reducing overwhelmingly complex data into only a few of the most useful dimensions. Once the strongest principal components have been identified, researchers typically analyse the dimensions in order to understand their nature, often assigning them descriptive names and determining the characteristic that seem to be most typical of each end of both vectors. Like Cluster Analysis, Principal Component Analysis can be carried out with most statistical packages.

5 Findings

5.1 Overall frequencies of AWs

The frequencies of academic words were calculated for each combination of genre and discipline, giving six subsets of data. The frequencies and distributions are given in Figure 3 and Table 1.

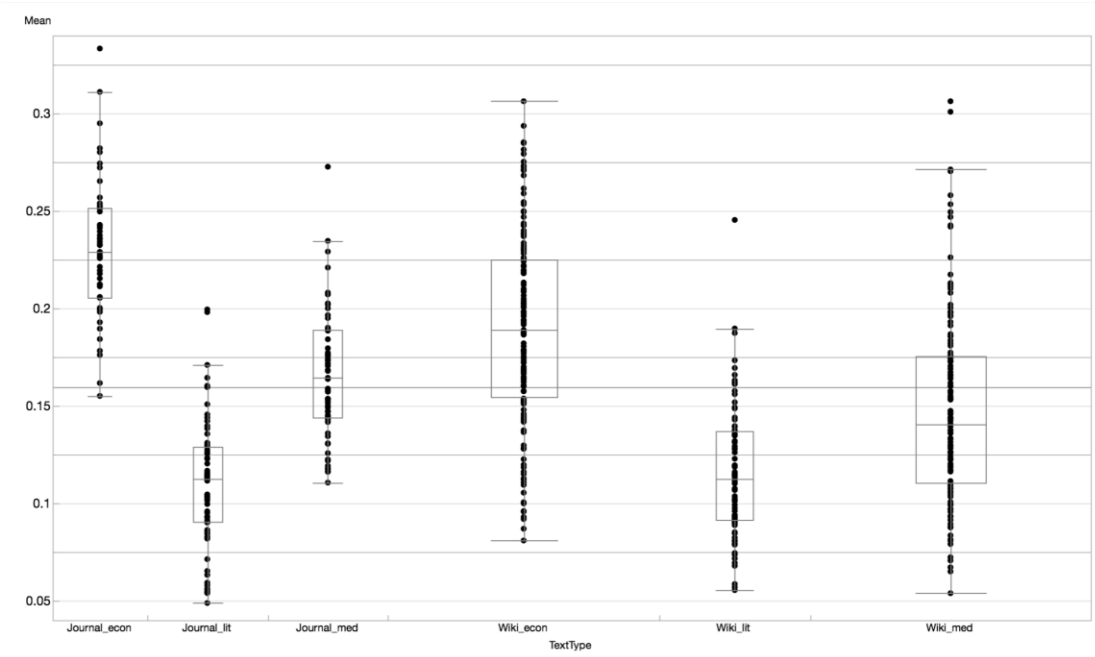


Figure 3. Box-and-whiskers plot of academic word frequencies in the six subcorpora

Table 1. Descriptive statistics of AW frequencies across subcorpora.

Level	n	Mean normalised frequency (per 1,000 words)	Std Dev	Lower 95%	Upper 95%
Journal_econ	50	0.230397	0.035985	0.22017	0.24062
Journal_lit	64	0.110057	0.033417	0.10171	0.1184
Journal_med	64	0.165722	0.032795	0.15753	0.17391
Wiki_econ	470	0.212721	0.070928	0.20629	0.21915

Wiki_lit	182	0.120429	0.042822	0.11417	0.12669
Wiki_med	439	0.14785	0.061209	0.14211	0.15359

As shown in the figure, both genres (RAs and Wikipedia articles) display similar patterns of variation between the three disciplines: economics has the highest mean frequency of AWs, followed by medicine and literary criticism. A non-parametric Mann-Whitney U-test for significance was carried out to test each pair.[14] With the exception of literature research articles and wiki articles, each pair was found to show a statistically significant difference (Table 2).

Table 2. Pairwise comparisons between the subcorpora (Mann-Whitney-Wilcoxon)

Level 1	Level 2	Z	p-Value
Wiki_econ	Journal_lit	11.0475	***
Wiki_econ	Journal_med	5.8145	***
Wiki_med	Journal_lit	5.0433	***
Wiki_med	Wiki_lit	5.3999	***
Journal_med	Journal_lit	7.6416	***
Wiki_lit	Journal_lit	1.4306	0.1525
Journal_med	Journal_econ	-7.5463	***
Journal_lit	Journal_econ	-8.9967	***
Wiki_econ	Journal_econ	-2.9279	***
Wiki_med	Journal_med	-3.5778	***
Wiki_lit	Journal_med	-7.3961	***
Wiki_lit	Journal_econ	-10.169	***

[14] The Mann-Whitney U-test is also known as the Wilcoxon test. It is a rank-sum test commonly used as a non-parametric equivalent to student's t-test. Non-parametric tests for significance are generally more appropriate for inferential analysis of linguistic data because they make no a priori assumptions about probability distributions.

Wiki_med	Journal_econ	-8.7982	***
Wiki_med	Wiki_econ	-14.054	***
Wiki_lit	Wiki_econ	-15.2162	***

This pattern is confirmed by the analysis of the dispersion of academic words using the chunking method discussed in section 4.1. Table 3 gives the number of chunks in each subcorpus, the mean frequency of AWs per 200-word chunk, and the standard deviation. As the table shows, chunks of the same discipline appear remarkably similar regardless of genre. The Wikipedia chunks shows slightly lower frequencies on average, but at the same time higher standard deviations, which suggests that the chunks are less consistent when it comes to the use of academic words. This observation is consistent with the composition of the Wikipedia subcorpora, in which the texts range considerably in both length and topic matter.

Table 3. Distribution of AWs across 200-word chunks.

Genre	Discipline	N	Mean freq / chunk	Sd
RA	Economy	2,770	24.19	7.63
	Literary analysis	2,588	11.94	5.83
	Medicine	1,206	16.82	6.23
Wiki	Economy	4,042	19.99	8.58
	Literary analysis	2,275	12.08	5.88
	Medicine	3,968	15.22	7.39

5.2 Similarities in texts and words

Although the frequency-and-dispersion-based methods reported in section 5.1 demonstrated that there are statistically significant differences between the different genres and disciplines, they also leave many unanswered questions. The main shortcoming here is that the frequencies reported are the mean frequencies of all academic words per text or chunk, which means that although we get a general sense of which genres and disciplines have more or fewer academic words, we know nothing about the distributions of the individual words nor, more importantly, about the distributional properties of all the 570 word families taken together.[15] Although we can make the educated guess that there must be words that are used in certain genres or disciplines and not in others, unless we examine the frequencies of all the words individually, traditional corpus linguistic methods leave us in the dark about such distributional differences. By using the computational grouping methods described in section 4.2., we can take into account the frequencies of all the different word families at the same time, finding texts that show similar distribution profiles and, conversely, finding words that have a tendency to occur together and those that do not.

We begin with a cluster analysis of the texts based on word family frequencies. The results can be seen in the dendrogram given in Figure 4, to which we have added a legend to indicate the predominant text category in each of the clusters. Starting from the right-hand side, we see a cluster of economy RAs, then a cluster of medical RAs and closely related to that a cluster of medical Wikipedia articles, and so on. With the exception of the left-most cluster, which contains a variety of all types of Wikipedia articles, it is immediately apparent that the texts from the same subcorpus appear to cluster together. This is particularly clear when it comes to the RAs, and this is of course to be expected: unlike Wikipedia articles, academic research articles are all written by professional researchers, who are familiar with the disciplinary requirements and whose submissions are reviewed by editors, referees and copy editors before publication.

The clustering shows that if we look at what AWs are used in the texts and how often, medical and economy research articles are collectively very similar to each other, forming one of the major cluster in the dendrogram along with a small selection of medical Wikipedia articles.[16] Similarly, Wikipedia articles on

---

[15] To be more precise, the frequencies are the frequencies of word families, which subsume varying numbers of unique lexical items.

[16] The annotations give the genre and discipline of the vast majority of text in each cluster, but it should to be noted that each cluster may include a small number of texts from other genres or disciplines.

economy and medicine form another, research articles and Wikipedia articles on literary scholarship form a third cluster, and the final cluster, which shows a distinct lack of academic vocabulary, comprises various shorter Wikipedia articles representing all three subcorpora. Some of the articles in this final cluster are overview articles, which merely list and link to important topics (e.g. Outline of literature), which understandably offer few opportunities for using ALW. Other articles have yet not received extensive attention by Wikipedia editors, perhaps reflecting their marginal status in the field, and do not provide a full coverage of their topic; these deal with e.g. individual sub-disciplinary specialisms (e.g. Neoclassical synthesis, Oligopsony, Medical geology, Semiotic literary criticism).

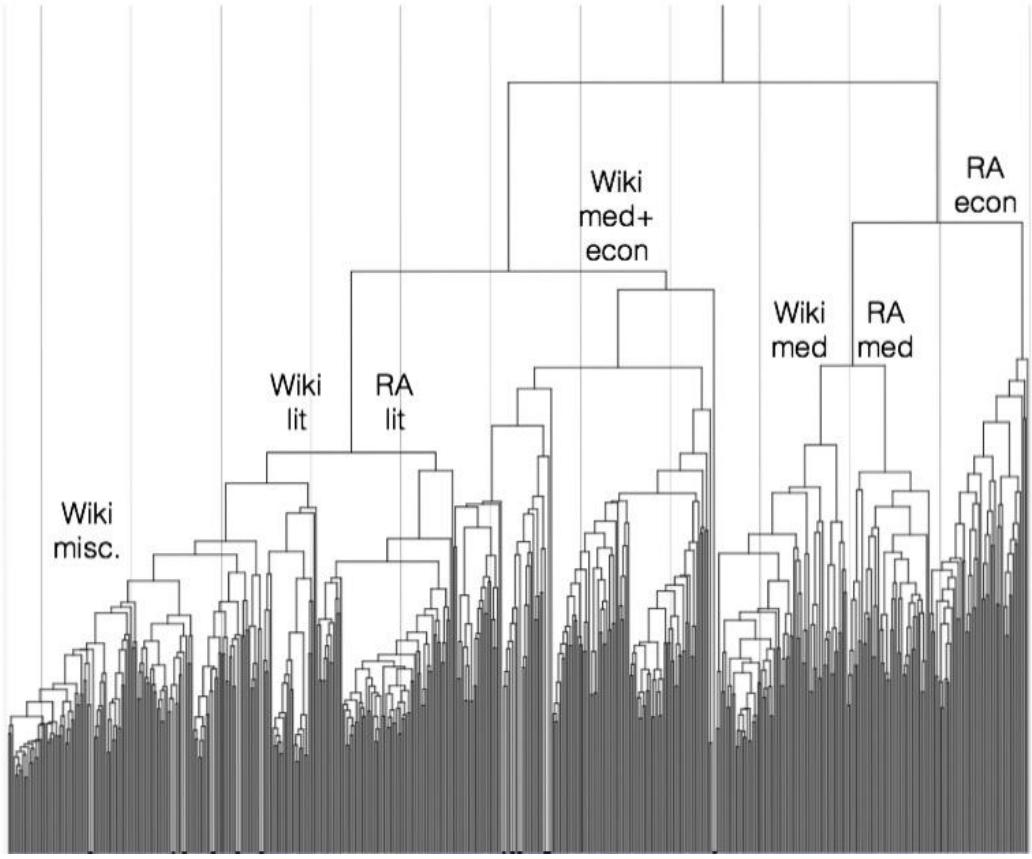


Figure 4. Clustering of articles based on the occurrence of AWs (Ward's method).

Consulting the output of cluster analysis, it can be observed that the clustering is partly explained by the presence of such AWs as medical and philosophy. In principle, these words may of course be used in a general sense in texts of different disciplines (e.g. in phrases like philosophy of science and philosophy of mind), but in our data they are almost exclusively found in just one of them.

However, the number of such words is too low to explain the clustering completely. Instead we must turn to disciplinary and stylistic similarities between the sub-corpora. What characteristics do research articles on economy and medicine share? What makes literary texts so different from both economy and medicine?

To answer the question in a data-driven fashion, we carried out a Principal Component Analysis (see Figure 5). Texts shorter than 2,000 words were left out of the analysis to mitigate the impact of low word counts on standardised frequency. Looking at the dispersion of lexical items, we can tentatively label the two main components as “text vs. data” (component 1, horizontal axis) and “theory vs. practice” (component 2, vertical axis). Component 1 explains 1.15% of the total variance, while component 2 explains 1.08%. Although these proportional contributions may seem small, it is worth observing that the full model includes 590 components.

Starting with the left-hand side of component 1, we see words related to textual topics such as text, author, publish, edit, lecture and comment. To the right, we see terminology related to data-driven scholarship such as vary, data, indicate, significant, outcome and hypothesis. It is not very difficult to see that there would be a significant difference in the distribution of these words across the disciplines: the former group would be associated with literary studies, where knowledge-making practices are typically intepretative and reiterative (see Groom 2009), and the latter with medicine and economics, which tend to rely on the application of agreed-upon quantitative methods for creating new knowledge (see e.g. Hyland 2000). Similarly, component 2 shows a distinct difference between the top and bottom half of the plot. Words found at the top of the plot are related to more theory-driven scholarship, such as undergo, proceed, incidence and confirm, while words at the bottom are relevant to the practical work, such as assess, detect, motivate, context and perspective.

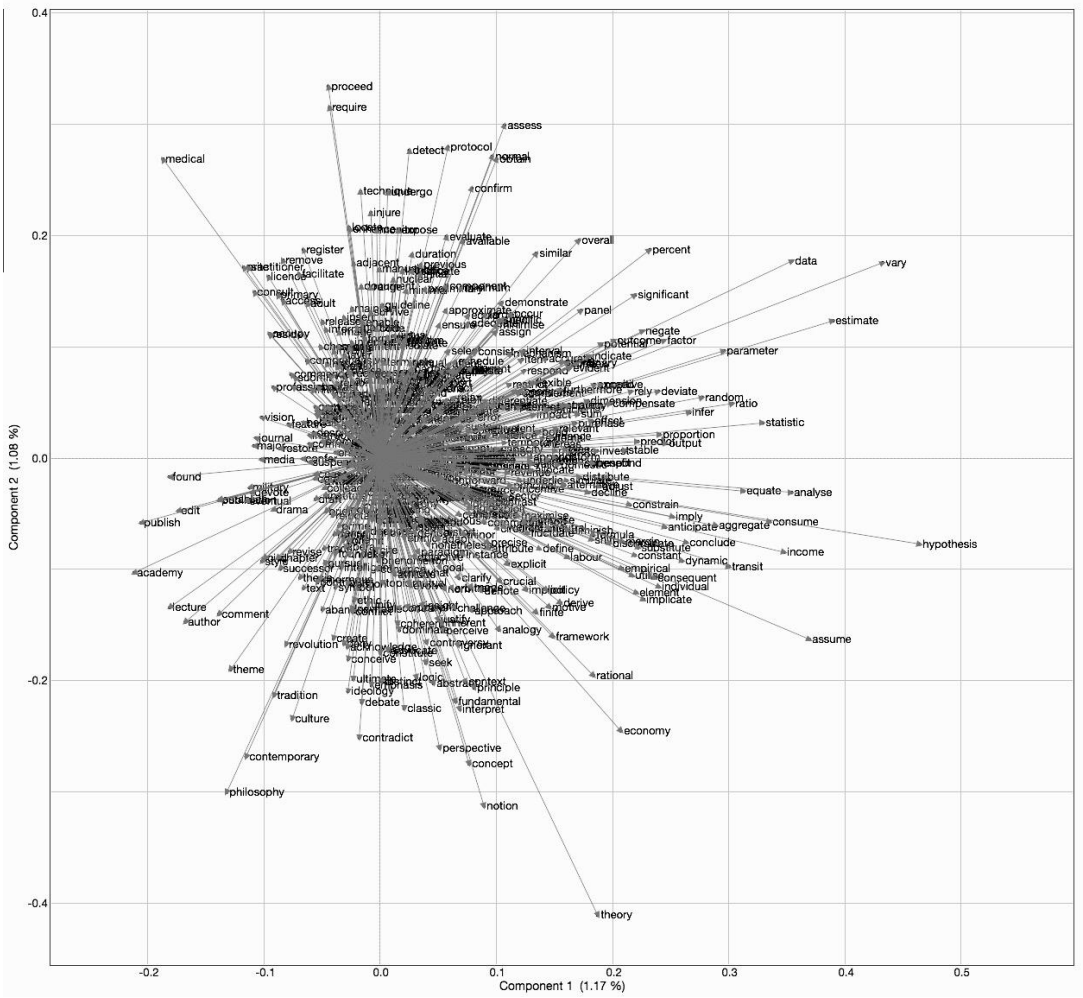


Figure 5. Principal component analysis of AWs.

The figure thus illustrates in very real terms how the co-variances between lexical items are related to differences in the distribution of disciplinary vocabulary, which in turn can be linked to well-known differences in the nature of disciplinary knowledge (Becher & Trowler 2001).

6 Discussion

Wikipedia articles have received numerous criticisms from academics, and in most cases these collaboratively edited encyclopaedia texts are obviously poor substitutes for research articles and textbooks authored by professional scientists and scholars. However, our analysis suggests that as far as vocabulary use is concerned, Wikipedia articles are not entirely different from RAs: AWs are frequently used in Wikipedia articles, too. The frequencies of AWs vary considerably between individual Wikipedia texts, which in part reflects the fact

that Wikipedia tend to be shorter, but such variation is also present in research articles. Our findings indicate that in both genre categories, discipline plays an important role in accounting for the observed variation in the frequency of academic words. AWs are by far the most commonly used in economics writing, while the lowest frequencies are found in literary criticism.

The findings reported in section 5 suggest that Wikipedia articles fall roughly into two major categories when it comes to the use of academic words. In the first category, we have long and detailed articles, which appear very similar to academic writing of the same discipline when it comes to the use of academic words. These articles tend to be on topics that are important, noteworthy and central. Examples of such articles include Electroencephalography, Free market and Lyric poetry. Consulting the article histories on the Wikipedia, we can observe that these articles have typically been frequently edited over several years, often by contributors who have expertise on the topic. The second category contains, shorter articles, which are less similar to academic research articles of the same discipline. These articles represent more niche topics, which consequently have been less incessantly edited. Examples of these articles include Gender-based medicine, Oligospony and Semiotic literary criticism

It should be emphasised that our findings do not carry value judgements. AW frequencies do not directly tell anything about the information density of texts: we may find comparatively low rates of AWs in passages with high rates of technical vocabulary. Individual texts may also display a remarkable amount of lexical variation even within a specific discourse domain, as demonstrated by Miller and Biber (2015) . Likewise, the use of AWs does not guarantee that they are used correctly, nor that the style of argumentation, the framing of questions or the rhetorical structure of the overall text would meet the standards of what competent members of the academic community would expect from a published text. And finally, it is important to keep in mind that the Academic Word List is primarily designed to represent the vocabulary of what might be described as generic academic language: words associated with scholarly argumentation, hedging and evaluation, among others. This means that discipline-specific words, such as Latin names of body parts, chemical compounds and mathematical terms are not included. Consequently, when we report that the mean frequency of AWs is lower in medical research articles than in economy research articles, we are only talking about items on the AWL, fully aware of the fact that the medical articles are very likely to include many items that simply do not show up in this analysis.

Our exploratory method is also well-suited to detecting general trends in multiple texts. With the Wikipedia texts in particular, the diverse and previously uncharted nature of the source texts means that the primary data ought to be approached with as few preconceived notions as possible. Thus, rather than focusing on the characteristics of individual texts or the use of individual words one by one, we consider it more useful to tackle the entire dataset as a whole and to approach it in a data-driven fashion, letting the distributional patterns

determine the noteworthy similarities and dissimilarities. It is worth remembering here that although we discussed the findings on the level of the 570 word families in Coxhead's academic word list for reasons of simplicity, the analysis involved roughly 3,000 word types. Without computational grouping methods, it would be virtually impossible to form a coherent picture of the distribution patterns of such a large lexical field in a data set that spans more than 1,200 texts and some 3.4 million words. The two first components identified using Principal Component Analysis, which we tentatively named "text vs. data" and "theory vs. practice", also have great intuitive appeal.

In sum, dismissing Wikipedia's style out of hand as non-academic is unnecessary. Much of Wikipedia may already be comparable to academic prose in terms of vocabulary use, and the quality of articles is likely to improve over time, especially as the encyclopaedia is increasingly reaching out to academics for help in editing the articles (see e.g. Hodson 2015, Schulenberg 2016). What is not going to change is the communicative purpose of Wikipedia, which is fundamentally different to that of research genres, and therefore use of Wikipedia in higher education will probably remain a source of some contention. We suggest that it is these generic differences, and the concomitant differences in argumentation styles, that are central to determining Wikipedia's role and function in educational settings, and EAP instruction should certainly attempt to raise students' awareness of these issues.

References

Barton, Matt & Robert E. Cummings eds. 2008. Wiki writing: collaborative learning in the college classroom. Ann Arbor: University of Michigan Press.

Biber, Douglas 1993. Representativeness in corpus design. Literary and linguistic computing 8:4, 243-257.

Biber, Douglas and James K. Jones (2009). Quantitative methods in corpus linguistics. In: Corpus Linguistics. An International Handbook. Volume 1. Ed. by Anke Lüdeling and Merja Kytö. Berlin: Mouton de Gruyter, 1286–1304.

Brezina, Vaclav and Dana Gablasova. 2013. Is there a core general vocabulary? Introducing the New General Service List. Applied Linguistics 31(1): 1–22.

Coxhead, Averil. 2000. A New Academic Word List. TESOL Quarterly, 34(2): 213–238.

Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. Proceedings of WWW. Available at: http://www.mpi-sws.org/~cristian/Echoes_of_power_files/wikipedia.talkpages.README.v1.01.txt.

Evert, Stefan 2006. How Random is a Corpus? The Library Metaphor. Zeitscrift für Anglistik und Amerikanistik 52:2, 177-190.

Gardner, Dee and Mark. Davies 2014. A New Academic Vocabulary List. Applied Linguistics 35(3): 305–327.

Gries, Stefan. 2008. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13:4, 403–37.

Gries, Stefan 2009. Dispersions and adjusted frequencies in corpora: Further explorations. In: Language, People and Numbers, ed. by Stefan Th. Gries, Stefanie Wulff and Mark Davies. Amsterdam: Rodopi. 197-212.

Groom, Nicholas 2009. Phraseology and epistemology in academic book reviews: a Corpus-Driven Analysis of Two Humanities Disciplines. In: Academic Evaluation. Review Genres in University Settings, ed. by Ken Hyland and Giuliana Diani. London: Palgrave Macmillan. 122–139.

Hiebert, Elfrieda. H. and Shira Lubliner. 2008. The nature, learning, and instruction of general academic vocabulary. In: What Research Has to Say About Vocabulary Instruction, ed. by Alan E. Farstrup and S. Jay Samuels. Newark: International Reading Association. 106–29.

Hiltunen, Turo. 2014. Choice of national variety in the English-language Wikipedia. In. J. Tyrkkö and S. Leppänen (eds.) Texts and discourses of the new media. Studies in Variation, Contacts and Change in English 15. Helsinki: Research Unit for Variation, Contacts, and Change in English. Available at: http://www.helsinki.fi/varieng/series/volumes/15/hiltunen/

Hodson, Richard. 2015. Wikipedians reach out to academics. Nature, 7 September 2015. url: http://www.nature.com/news/wikipedians-reach-out-to-academics-1.18313.

Hoover, David L. 2003. Multivariate analysis and the study of style variation. Literary and Linguistic Computing 18(4). 341–360.

Hyland, Ken and Polly Tse. 2007. “Is there an ‘academic’ vocabulary”. TESOL Quarterly 41:2, 235–253.

Ide, Nancy et al. 2002. The American National Corpus. Proceedings of the 3rd Language Resources and Evaluation Conference LREC, Canary Islands. Paris: ELRA.

Kamm, Oliver. 2007. “Wisdom? More like dumbness of the crowds”. The Sunday Times, 16 August 2007.

Kilgarriff, Adam. 2005. Language is never, ever, ever random. Corpus Linguistics and Linguistic Theory 1(2), 263–276.

Kuteeva, Maria. 2001. Wikis and academic writing: Changing the writer–reader relationship. English for Specific Purposes 30(1): 44–57.

Manning, Christopher and Hinrich Schütze 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Martinez, Ron and Norbert Schmitt. 2015. “Vocabulary”. In: The Cambridge handbook of English corpus linguistics, ed. by. Douglas Biber and Randi Reppen, 439–459. Cambridge: Cambridge University Press.

Medelyan, Olena, David Milne, Catherine Legg and Ian H. Witten. 2009. Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67(9): 716–754.

Miller, Don and Douglas Biber. 2015. Evaluating reliability in quantitative vocabulary studies. The influence of corpus design and composition. International Journal of Corpus Linguistics 20(1): 30–53.

Miller, Julia. 2012. Building academic literacy and research skills by contributing to Wikipedia: A case study at an Australian university. Journal of Academic Language and Learning 8(2): A72–A86.

Nation, I. S. P. 2001. Learning Vocabulary in Another Language. Cambridge: CUP.

Rosenzweig, Roy 2006. Can History be Open Source? Wikipedia and the Future of the Past. The Journal of American History 93(1): 117-46.

Schulenberg, Frank. 2016. The Wikipedia Year of Science is here! Wiki Education Foundation, 19 January 2016. url: https://wikiedu.org/blog/2016/01/19/wikipedia-year-of-science/

Shaoul, Cyrus & Westbury Chris. 2010. The Westbury Lab Wikipedia Corpus. Edmonton, AB: University of Alberta. (downloaded from [http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html]

Swales, John M. 1990. Genre Analysis. English in Research Settings. Cambridge: Cambridge University Press.

Tardy, Christine M. 2010. Writing for the world: Wikipedia as an introduction to academic writing. English Teaching Forum 48(1): 12–19.

West, Michael. 1953. A General Sercive List Of English Words. London: Longman.

Waters, Neil L. 2007. "Why you can't cite Wikipedia in my class. Communications of the ACM 50(9): 15–17.

Yelland, Michael, Paul P. Glasziou, Nikolai Bogduk, Philip J. Schluter, and Mary McKernon. 2004. Prolotherapy Injections, Saline Injections, and Exercises for Chronic Low-Back Pain: A Randomized Trial. Spine 29: 9-16.